수집된 데이터 및 데이터 전처리 문서

전처리 결과 요약			
데이터 소스	데이터 수량	**json** 파일 크기	**Faiss** 벡터 파일 크기
네이버 크롤링 데이터	22877 개	63.7 MB	252.6 MB
오픈 소스 데이터	3478 개	1.8 MB	23.5 MB
총 합	26355 개	65.5 MB	276.1 MB

1. 수집 데이터

- 수집 방법: 크롤링, 오픈 데이터 소스
- 수집 기간: 12/23 - 1.3
- 도구 및 기술: Selenium, BeautifulSoup
- 데이터 소스: 네이버지도, 공공데이터포탈 (서울관광재단), LOCALDATA

- 네이버 지도 크롤링 데이터

크롤링 키워드	음식점, 술집, 카페	동 단위 총 180개 키워드 검색 예시) 강남구 논현동 카페
총 데이터 수	47700개	
컬럼 수	17 개	store_name,category,new_open,rating,visited_review,directions_text,store_id,address,blog_review,phone_num,business_hours,info,convenience_facilities_and_services,Parking,sns,review_like_this_part_output,menu
형식	csv	

- 오픈 소스 데이터

데이터 이름	사이트 이름	특징	링크	파일크기	갱신 년월	데이터 수	컬럼
한국보건산업진흥원_외국인환자유치기관현황	공공데이터포탈	한국보건산업진흥원 외국인환자 유치기관 현황 자료입니다. (유치기관 상태, 유치기관명, 유치기관유형, 지역, 유치대상국가 등)	https://www.data.go.kr/data/3050000/fileData.do	1.5 MB	2024-07	1738개	상호,대표자,기관구분,주소,타겟국가
서울시 의료관광허가 의료기관 정보 (한국어) 의료관광허가가된 의료기관의 상호 및 주소, 타겟국가 등 관련 정보 제공(한국어)	공공데이터포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-12973/S/1/datasetView.do	226 KB	2024-04		
서울시 관광음식 (서울의 맛집 탐방. 한식, 중식, 일식, 아시아식, 서양식, 주점 등 메뉴별 맛집, 카페&디저트, 채식, 할랄 등 특화	공공데이터포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21054/S/1/datasetView.do	453 KB	매월 4일	725개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,교통정보,홈페이지 언어,대표메뉴

식당까지 정보 제공)							
서울시 관광명소 (서울의 랜드마크, 고궁, 역사적 장소, 오래가게, 미술관, 박물관 등 놓칠 수 없는 서울의 명소 장소 정보)	공공데이터 포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21050/S/1/datasetView.do	219 KB	매월 4일	280개	상호명,콘텐츠URL,주소,신주소,전화번호,팩스번호,웹사이트,운영시간,운영요일,휴무일,교통정보,태그,장애인편의시설
서울특별시_관광 쇼핑 (서울의 쇼핑몰, 백화점, 면세점, 마켓, 뷰티샵, 한류 및 관광상품 등을 세계 각종 언어로 상호명, 주소 위치정보 등 공개합니다.)	공공데이터 포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21053/S/1/datasetView.do	94 KB	매월 4일	153개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,교통정보
서울특별시_관광 문화 (주요 언어로 관광시설,문화시설, 휴식시설, 스포츠시설, 놀이공원, 테마파크, 체험공간 등 익사이팅한 서울의 놀거리들	공공데이터 포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21052/S/1/datasetView.do	116 KB	매월 4일	13개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,운영요일,휴무일,교통정보,홈페이지 운영 언어,유모

정보를 안내)							차 대여 여부
서울특별시_관광 자연(도심 속에서 자연의 힐링을 느낄 수 있는 서울의 산, 강, 계곡, 섬, 공원, 정원, 식물원 등 자연 관광명소들을 언어별로 소개합니다.)	공공데이터 포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21051/S/1/datasetView.do	56 KB	매월 4일	33개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,운영요일,휴무일,교통정보
서울특별시_관광거리 정보 (서울특별시의 주요 관광거리에 대한 공식 명칭 및 주소 등 정보를 제공합니다.(한국어) 검색키워드,최종표기명,지번주소,법정시,법정구,법정동,위치정보 등을 제공합니다.)	공공데이터 포탈 (서울관광재단)	경도 위도 포함	https://data.seoul.go.kr/dataList/OA-12929/S/1/datasetView.do	25 KB	2016-02	54개	최종 표기명,지번 주소,중심 좌표 X,중심 좌표 Y
서울특별시_관광안내소 (표준 데이터) 서울특별시 관광안내소 정보(안내	공공데이터 포탈 (서울관광재단)	경도 위도 포함	https://data.seoul.go.kr/dataList/OA-20350/S/1/datasetView.do	7 KB	2024-12	8개	관광안내소명,안내소위치명,안내소소개,부가서비스정보,

소소개, 부가서비스 정보, 휴무일, 운영시간, 근무인원수, 외국어안내 가능여부, 전화번호 등)를 제공합니다							휴무일,운영시작시각(하절기),운영종료시각(하절기),운영시작시각(동절기),운영종료시각(동절기),안내가능외국어,안내소전화번호,소재지도로명주소,소재지지번주소,홈페이지주소,위도,경도
관광숙박업	LOCALDATA	경도 위도 데이터 포함	https://www.localdata.go.kr/data/dataView.do	186 KB		474개	인허가일자,소재지전화,소재지전체주소,도로명전체주소,도로명우편번호,사업장명,좌표정보X(EPSG5174),좌표정보Y(EPSG5174),관광숙박업상세명

2. 전처리

네이버 지도

- 데이터 병합 **:** 수집한 데이터를 한번에 처리하기 위해 각 키워드별 데이터 병합
    - 행을 기준으로 하여 병합
        - pandas 라이브러리의 concat() 함수 사용
    - 총 180개의 데이터 병합
        - 행정동 단위 키워드
            - ex) 강남구 논현동 카페, 강남구 역삼동 카페
- 중복 제거
    - store_id 컬럼을 기준으로 drop_duplicates() 함수를 이용하여 중복 제거
- 결측치 처리
    - 네이버지도 데이터 23693개 중 store_name 컬럼 816개 결측
        - 3.4% 결측치 존재
    - pandas 라이브러리의 dropna() 함수를 통해 결측치 제거
- 데이터 분리
    - 강남구, 용산구, 종로구, 중구로 구별 필터링 하여 파일 분리
- 데이터 변환
    - csv 파일에서 json 형식으로 변환
        - `to_json.()`함수를 사용하여 변환
- 벡터 스토어 임베딩
    - faiss 라이브러리를 사용해서 json 형식의 데이터를 벡터 스토어로 임베딩
    - 벡터 스토어 구조

meta data	page content
store_name, store_id, address	store_name, category, new_open, rating, visited_review, directions_text, address, blog_review, phone_num, business_hours, info, convenience_facilities_and_services, Parking, sns, review_like_this_part_output,menus

오픈 소스 데이터

- 행정구 단위 데이터 분리
    - 9개의 데이터를 각각 강남구, 종로구, 중구, 용산구 지역별로 데이터 필터링 하여 파일 분리,
        - 파일 수 : 9개 → 36개
        - 9개의 데이터 : 음식점, 관광거리, 명소, 외국인 유치 병원, 관광 안내소, 문화 시설, 자연 체험, 숙박업, 쇼핑거리
- 데이터 변환
    - csv 파일에서 json 형식으로 변환

        - to_json.()함수를 사용하여 변환
- 행정구 별 데이터 병합
    - 36개의 데이터 강남구, 종로구, 중구, 용산구 지역별로 파일 병합
        - 파일 수 : 36개 → 4개
- 벡터 스토어 임베딩
    - faiss 라이브러리를 사용해서 json 형식의 데이터를 벡터 스토어로 임베딩
    - 벡터 데이터 구조

데이터	**metadata**	**page_content**내용
음식점	상호명, 주소	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 교통정보, 홈페이지 언어, 대표메뉴
관광 거리	최종 표기명 , 지번 주소	최종 표기명, 지번 주소
명소	상호명, 주소	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 팩스번호, 웹사이트, 운영시간, 운영요일, 휴무일, 교통정보, 태그, 장애인편의시설
관광 안내소	관광안내소명, 소재지 지번주소	관광안내소명, 안내소위치명, 안내소소개, 부가서비스정보, 휴무일 ,운영시작시각(하절기) , 운영종료시각(하절기), 운영시작시각(동절기), 운영종료시각(동절기), 안내가능외국어, 안내소전화번호 ,소재지도로명주소, 소재지지번주소, 페이지주소
문화 시설	상호명, 주소	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 운영요일, 휴무일, 교통정보, 홈페이지 운영 언어, 유모차 대여 여부
외국인 유치 병원	상호, 주소	상호, 대표자, 기관구분, 주소, 타겟국가

쇼핑센터(거리)	상호명, 주소	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 교통정보
자연체험	상호명, 주소	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 운영요일, 휴무일 , 교통정보
숙박업	사업장명, 소재지전체주소	인허가일자, 소재지전화, 소재지전체주소, 도로명전체주소 , 도로명우편번호, 사업장명, 관광숙박업상세명

3. 부록

1. json 파일 예시

a. 네이버 지도 json 파일


```
{

"Unnamed: 0":1,
"store_name":"윈스커피",
"category":"카페,디저트",
"new_open":"",
"rating":"별점",
"visited_review":"방문자 리뷰 557",
"directions_text":"삼성역 5번 출구에서 400m 직진 후 NC타워 차움검진센터에서 우회전 후 90m앞에 위치한 윈스커피입니다.\n\n또한 코엑스 도심공항터미널 쪽으로 나오셔서 4분정도만 걸어오시면 분위기 좋은 삼성동 로스터리 카페 윈스커피에 도착하실 수 있습니다:)\n내용 더보기",
"store_id":"1372135556",
"address":"서울 강남구 테헤란로83길 18 1층",
"blog_review":"블로그 리뷰 421",
"phone_num":"070-8863-0763",
"business_hours":"월\n08:00 - 19:30\n화\n08:00 - 19:30\n수(1/1)\n새해 첫날 휴무\n목\n08:00 - 19:30\n금\n08:00 - 19:30\n토\n10:00 - 19:00\n일\n정기휴무 (매주 일요일)\n",
"info":"삼성동 코엑스 바로 옆에 위치한 로스터리 윈스커피에서는 직접 선별한 높은 퀄리티의 생두를 로스팅한 3가지 블랜드 및 디카페인 커피를 즐기실 수 있습니다.\n\n윈스커피는 좋은 품질의 원재료를 사용한 시그니처 음료들과 건강하고
```

신선한 재료로 만든 그릭요거트 그리고 프랑스 프리미엄버터를 사용해 만든 윈스 크로플을 선보입니다.\n\n- 윈스커피에서 만날 수 있는 특별한 메뉴 -\n>직접 로스팅한 3가지 다른 컨셉의 블랜드와 디카페인 커피\n>마다가스카르산 최고의 바닐라빈으로 만들어 깊고 진한 달콤함이 돋보이는 바닐라빈 라떼\n>직접 만든 땅콩 크림이 라떼 위에 올려져 고소하고 달달한 피넛슈페너\n>프랑스 프리미엄 버터를 사용한 시그니처 윈스 크로플\n>꾸덕꾸덕하고 고소한 요거트에 수제 그래놀라와 신선한 과일이 듬뿍 올라간 그릭요거트\n\n모던하고 감성적인 분위기로 꾸며진 윈스커피에서 맛있는 음료와 함께 행복한 시간을 보내세요:)\n\n삼성역카페 윈스커피는 코엑스 도심공항에서 걸어서 4분거리에 위치해 있습니다.\n펼쳐보기",
"convenience_facilities_and_services":"로스터리\n유기농 메뉴\n포장\n남/녀 화장실 구분\n단체 이용 가능\n배달\n무선 인터넷\n",
"Parking":"주차 정보없음",
"sns":"https://smartstore.naver.com/wynscoffee\n",
"review_like_this_part_output":"\"커피가 맛있어요\"이 키워드를 선택한 인원284\n\"음료가 맛있어요\"이 키워드를 선택한 인원165\n\"디저트가 맛있어요\"이 키워드를 선택한 인원112\n\"인테리어가 멋져요\"이 키워드를 선택한 인원93\n\"친절해요\"이 키워드를 선택한 인원90\n\"대화하기 좋아요\"이 키워드를 선택한 인원73\n\"매장이 청결해요\"이 키워드를 선택한 인원48\n\"가성비가 좋아요\"이 키워드를 선택한 인원44\n\"특별한 메뉴가 있어요\"이 키워드를 선택한 인원41\n\"사진이 잘 나와요\"이 키워드를 선택한 인원23\n\"집중하기 좋아요\"이 키워드를 선택한 인원20\n\"아늑해요\"이 키워드를 선택한 인원13\n\"좌석이 편해요\"이 키워드를 선택한 인원7\n\"매장이 넓어요\"이 키워드를 선택한 인원7\n\"화장실이 깨끗해요\"이 키워드를 선택한 인원6\n\"차분한 분위기에요\"이 키워드를 선택한 인원5\n\"음악이 좋아요\"이 키워드를 선택한 인원5\n\"뷰가 좋아요\"이 키워드를 선택한 인원5\n\"메뉴 구성이 알차요\"이 키워드를 선택한 인원4\n\"오래 머무르기 좋아요\"이 키워드를 선택한 인원3\n\"건강한 맛이에요\"이 키워드를 선택한 인원3\n\"양이 많아요\"이 키워드를 선택한 인원2\n\"음식이 맛있어요\"이 키워드를 선택한 인원1\n\"재료가 신선해요\"이 키워드를 선택한 인원1\n\"종류가 다양해요\"이 키워드를 선택한 인원1\n",
"menu":"아메리카노대표고소함과 깔끔함이 돋보이는 밸런스 좋은 커피5,500원\n수제 바닐라빈 라떼대표마다가스카르산 최고의 바닐라빈으로 직접만든 깊고 진한 달콤함7,000원\n윈스 크로플대표13,000원\n그릭요거트대표고소하고 상큼한 그릭요거트위에 수제 그래놀라와 신선한 과일이 가득9,000원\n피넛슈페너대표부드러운 수제 땅콩 크림과 팜트리 라떼8,000원\n카페라떼대표6,000원\n아인슈페너팜트리 블랜드 커피와 부드럽고 달콤한 크림7,000원\n말차라떼100% 제주 유기농 햇말차7,500원\n에스프레소5,000원\n카푸치노6,000원\n100% 생과일 ABC주스사과 당근 비트 조합으로 매장에서 직접 만든 상큼한 디톡스 주스7,000원\n문경 오미자 에이드100% 오미자 원액7,000원\n코코넛 스무디7,500원\n배슬러시6,000원\n블루베리요거트스무디7,000원\n"
}

b. 오픈 소스 데이터 json 파일

- 음식점

{
"Unnamed: 0":5,

"상호명":"개미집",
"콘텐츠URL":"https://korean.visitseoul.net/restaurants/개미집/KOP001095?utm_source=seoulopendata&utm_medium=restaurants&utm_content=KOP001095",
"주소":"135-888 서울 강남구 신사동 528-4 ",
"신주소":"06028 서울 강남구 압구정로 110 (화인빌딩) ",
"전화번호":"02-541-5955",
"웹사이트":"http://gaemizip.modoo.at/",
"운영시간":"16:30~24:00",
"교통정보":"3호선 신사역 8번 출구 ",
"홈페이지 언어":NaN,
"대표메뉴":"매운갈낙찜, 모짜렐라김치전, 홍어삼합, 제철사시미"
}

- 관광거리

{
"Unnamed: 0":1,
"최종 표기명":"명동거리",
"지번 주소":"서울시 중구 명동 일대",
"중심 좌표 X":126.9788194313,
"중심 좌표 Y":37.568058597
}

- 명소

{
"Unnamed: 0":0,
"상호명":"1898 명동성당 ",
"콘텐츠URL":"https://korean.visitseoul.net/attractions/1898-명동성당/KOP015338?utm_source=seoulopendata&utm_medium=attractions&utm_content=KOP015333",
"주소":"100-809 서울 중구 명동2가 1-1 ",
"신주소":"04537 서울 중구 명동길 74 (명동2가, 명동성당) ",
"전화번호":"02-774-1784",
"팩스번호":NaN,
"웹사이트":NaN,
"운영시간":"성당사무실 화 ~ 금 | 09:00 ~ 20:30 토 요 일 | 09:00 ~ 20:00 일 요 일 | 09:00 ~ 21:00 ",
"운영요일":NaN,
"휴무일":"설날, 추석 당일 (성당사무실: 월요일 휴무)",
"교통정보":"2,3호선 을지로 3가역 12번 출구에서 약 403m (도보 7분) 4호선 명동역 10번 출구에서 약 425m (도보 6분)",
"태그":" 성당,복합문화공간, 고딕, 1898광장,명동,명동대성당,명동나들이",
"장애인편의시설":NaN
}

- 문화시설

{
"Unnamed: 0":1,"상호명":"컬러풀뮤지엄(COLORPOOL MUSEUM)",

"콘텐츠URL":"https://korean.visitseoul.net/entertainment/컬러풀뮤지엄COLORPOOL-MUSEUM/KOP034412?utm_source=seoulopendata&utm_medium=entertainment&utm_content=KOP034412",
"주소":" 서울 종로구 관훈동 155-2 ",
"신주소":"03145 서울 종로구 인사동길 49 (관훈동) 안녕인사동 6층",
"전화번호":"02-6954-2872",
"웹사이트":"http://colorpoolmuseum.com/",
"운영시간":"10:00 ~ 21:00 (20:20 매표마감)",
"운영요일":"연중무휴",
"휴무일":NaN,
"교통정보":"3호선 안국역 6번 출구에서 약 180m (도보 2분) ",
"홈페이지 운영 언어":"한국어,영어,일어, 중문(번체), 중문(간체)",
"유모차 대여 여부":NaN
}

- 숙박업

{
"Unnamed: 0":1,
"인허가일자":"2024-12-19",
"소재지전화":NaN,
"소재지전체주소":"서울특별시 중구 남산동2가 32-13 ",
"도로명전체주소":"서울특별시 중구 퇴계로18길 33, 지하1층,지상4층 (남산동2가)",
"도로명우편번호":"04631","사업장명":"센트럴",
"좌표정보X(EPSG5174)":198623.369754854,
"좌표정보Y(EPSG5174)":450781.146728928,
"관광숙박업상세명":"호스텔업"
}

- 외국인 유치 병원 데이터

{
"상호":"(의) 성광의료재단 강남차의원",
"대표자":"차경섭",
"기관구분":"의원",
"주소":"서울특별시 강남구 역삼동 605",
"타겟국가":NaN
}

- 관광안내소 데이터

{
"관광안내소명":"강남관광정보센터",
"안내소위치명":"3호선 압구정역 6번 출구 도보 2분",
"안내소소개":"서울 및 강남 주요 관광명소・교통・음식・숙박・행사 관련 최신 정보를 영중일 다국어로 안내",
"부가서비스정보":"커뮤니케이션 공간+공유오피스+사무공간+체험존+휴게시설+로봇카페",
"휴무일":"토+일+공휴일",
"운영시작시각(하절기)":"10:00",
"운영종료시각(하절기)":"18:00",
"운영시작시각(동절기)":"10:00",

"운영종료시각(동절기)":"18:00",
"안내가능외국어":"영어, 일본어, 중국어, 러시아어",
"안내소전화번호":"02-1661-2230",
"소재지도로명주소":"서울특별시 강남구 압구정로 161, 1층",
"소재지지번주소":"서울특별시 강남구 압구정동 428, 1층",
"홈페이지주소":"https://www.gangnam.go.kr/contents/tourgangnam/2/view.do?mid=FM010406",
"위도":37.5268431,
"경도":127.0270587
}

- 쇼핑 데이터

{
"상호명":"롯데백화점 에비뉴엘관",
"콘텐츠URL":"https://korean.visitseoul.net/shopping/롯데백화점-에비뉴엘관/KOP009521?utm_source=seoulopendata&utm_medium=shopping&utm_content=KOP009521",
"주소":" 서울 중구 소공동 남대문로 2가 130",
"신주소":"04533 서울특별시 중구 남대문로 73 (남대문로2가) ",
"전화번호":"02-771-2500",
"웹사이트":"https://www.lotteshopping.com/store/main?cstrCd=0394",
"운영시간":"월~목 10:30 - 20:00 금요일, 주말 10:30 - 20:30",
"교통정보":"4호선 명동역 5번출구 도보 10 ~ 15분 2호선 을지로입구역 7,8번 출구 도보 5분 1호선 시청역 7번 출구 도보 10분"
}

- 자연 체험 데이터

{
"상호명":"가을단풍길 - 남산남측순환로",
"콘텐츠URL":"https://korean.visitseoul.net/nature/남산남측순환로/KOP036947?utm_source=seoulopendata&utm_medium=nature&utm_content=KOP036947",
"주소":" 서울 용산구 용산동2가 산 1-3 ",
"신주소":"04340 서울 용산구 남산공원길 126 (용산동2가) ",
"전화번호":"02-2199-7623(용산구청 공원녹지과)",
"웹사이트":"https://www.seoul.go.kr/storyw/autumn/list.do",
"운영시간":NaN,
"운영요일":NaN,
"휴무일":NaN,
"교통정보":"3, 4호선 충무로역 또는 3호선 동대입구역 순환버스 01번(남산순환) 탑승 후 '서울타워' 하차"
}